



Erstellung von synthetischen Daten zur Nutzung des Mikrozensus in der Hochschullehre

Amelie Plöger, Colin Schmidt, Sarah Redlich

IT.NRW – Statistisches Landesamt

23. Juni 2026

Campus Files in der Hochschullehre

- Statistische Methodenausbildung
- aktuelle realitätsnahe Daten
- niederschwellige und kostenfreie Bereitstellung
- Mikrozensusdaten mit moderatem Informationsverlust

Anforderungen

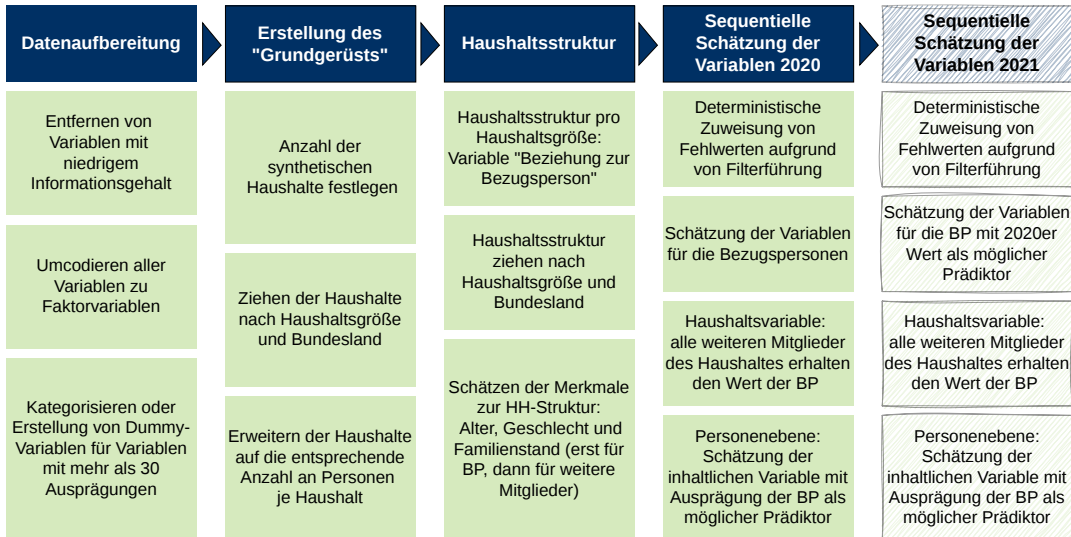
- Lerndatensatz: Forschungsergebnisse nur mit SUF oder On-Site-Daten
- uni- und multivariate Zusammenhänge
- metrische Merkmale z. B. für parametrische Verfahren
- Haushaltszusammensetzung
- später: Längsschnittverknüpfbarkeit
- keine Reidentifizierbarkeit von realen Personen und Haushalten

⇒ Lösung: synthetische Daten

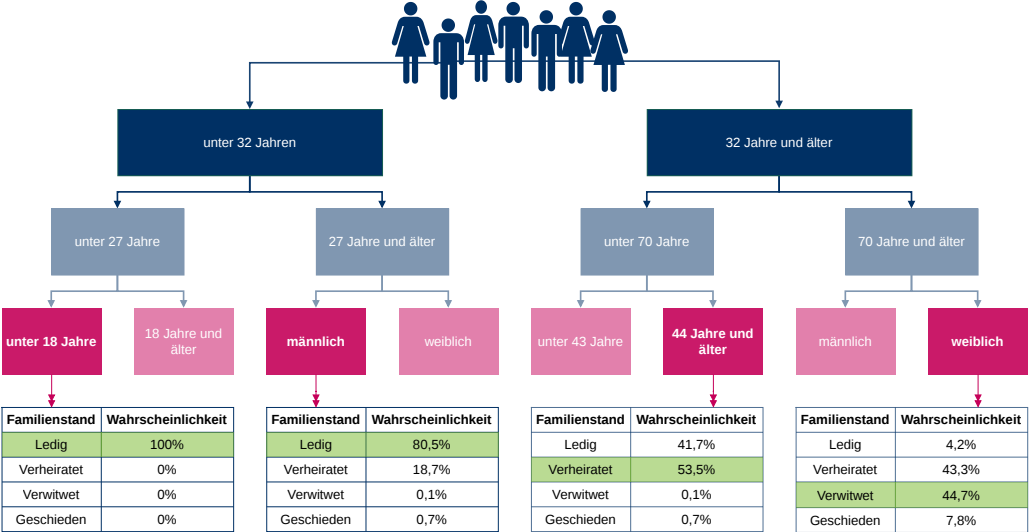
Datengrundlage

- Mikrozensus Scientific-Use-Files
- Pilotjahr 2020 (perspektivisch weitere Berichtsjahre)

Vorgehen



Fiktives Beispiel Entscheidungsbaum: Familienstand



Herausforderungen

Herausforderung	Lösung	Bewertung
hoher Rechenaufwand / lange Laufzeiten	Entfernen von Variablen mit niedrigem Informationsgehalt / R-Server	↻
komplexe Filterführung	deterministische Zuweisung von Fehlern	✓
Haushaltsstruktur	Behalten <u>einer</u> Originalvariable; Schätzung HH-Mitglieder unter Berücksichtigung der BP	✓
Einfluss der Variablenreihenfolge	Festlegen der Reihenfolge nach Wichtigkeit und Relevanz	↻
Variablen mit 30+ Kategorien	Zusammenfassen von Kategorien	✓
Klassifikationsvariablen mit 30+ Kategorien	Dummy-Codierung von hierarchischen Variablen	✓
weitere Herausforderungen	...	...

Nützlichkeit vs. Aufdeckungsrisiko

Nützlichkeit

- Ähnlichkeit der Randverteilungen → Hellinger Distanz
- Ununterscheidbarkeit originaler und synthetischer Daten → pMSE, pMSEratio

Aufdeckungsrisiko

- Vollständig synthetisch generierte Daten bergen praktisch kein Aufdeckungsrisiko
- Fälle des synthetischen Datensatzes sind nicht ähnlicher zu Trainingsdaten als zu Testdaten → Hamming Distanzen

Fazit und Ausblick

- ✓ Ähnlichkeit zu Originaldaten erzielt
- ✓ bisher **keine** überzufälligen Ähnlichkeiten zu Trainingsdaten
 - Methodenvergleiche und Festlegung von Parametern
 - Dokumentation und Abstimmung
 - Veröffentlichung (<https://campus-file-fdz.nrw.de/>)
 - regelmäßige Implementierung in die Aufbereitungsprozesse



Kontakt

Amelie Plöger und **Dr. Sarah Redlich**

S1-6 Übergreifende Analysen

Tel. +49211-9449-2901

uebergreifende-analysen@it.nrw.de

Colin Schmidt

S1-8.2 Forschungsdatenzentrum – Regionaler Standort

Tel. +49211-9449-2939

forschungsdatenzentrum@it.nrw.de